METHOD FOR SPECIFYING EQUIVALENCE OF LANGUAGE GRAMMARS AND AUTOMATICALLY TRANSLATING SENTENCES IN ONE LANGUAGE TO SENTENCES IN ANOTHER LANGUAGE IN A COMPUTER ENVIRONMENT

Field of Invention :

The invention relates to a method for specifying equivalence of language grammars and the automatic translation of sentences in one language to sentences in another language in a computer environment.

Background of the Invention :

A language is basically a set of sentences that can be formed by following certain rules. The basic building block of any language is its alphabet. There are numerous languages existing today in the same way. The sentences are a collection of words that are formed from the letters of the alphabet. There are certain rules to be followed when putting these words together. These rules are called grammar of the language and are unique for each and every language. These rules determine the valid sentences of the language. Thus one can define grammar as a concise specification using which, it is possible to generate all the valid sentences of the language. A grammar specifies the syntax or structure of a

language; irrespective of whether it is a language such as English or programming language such as 'C' or assembly language.

Very often, it is required to convert sentences in one language to equivalent sentences in another language. For example from English to French or from a programming language to assembly language. To perform such tasks the language grammars have to be specified and the source language statements should be validated and translated to sentences in the target language.

A method used in the prior art for translating a language to another language used is carried out in the following manner.

Define the source language grammar. Parse the sentences and convert them to a predefined intermediate format and translate finally, the intermediate format to the target language.

The disadvantages in performing a translation by the above mentioned method are the following.

(i) This method will not allow equivalence of the source language grammar and target language grammar to be specified. Thus there is

no real correspondence between source language grammar and target language grammar.

(ii) Normally this method allows the translation from one source language to one target language only. Mapping to multiple target languages will not be possible.

(iii) Mapping from a source language to a target language is predefined and thus supporting translations to new languages will be difficult.

Object of the Invention :

Bearing in mind the problems and detriments of the prior art, the object of the present invention is to provide a method to automatically translate sentences from one language to another, overcoming the above mentioned deficiencies.

Thus one of the object of the present invention is to be able to specify the equivalence of the source language grammar and target language grammar.

Another object of the present invention is to allow mapping to multiple target languages. Method according to the invention should have no restrictions to translating a source language to more than one target languages.

Description of the invention :

The invention provides a method for representing equivalence of language grammars and for the automatic translation of sentences in one language to sentences in another language in a computer environment.

Let L_1 to L_n be n number of languages and G_1 to G_n represent the respective grammars for the languages L_1 to L_n. Each grammar is unique to that particular language. Each grammar G_1 to G_n consists of a set of terminal symbols, a set of non-terminal symbols, a unique start symbol which is a nonterminal symbol and a set of production rules. These production rules are the main aspects of the grammar. Production rule define the rules to reduce a string of terminal and/or nonterminal symbols to a target nonterminal symbol.

In a grammar, there is at least one production rule that has the start symbol as its target nonterminal symbol. A sentence of a language may be defined as any string derived from the start symbol composed of only terminal symbols.

In the method according to the invention, a unified grammar specification is created for the grammars G_1 to G_n of all the languages L_1 to L_n respectively. Then the text in the source language is separated into a list of tokens using conventional lexical analyser for the source language. A nonterminal symbol is set to the start symbol of the unified grammar specification. Then a set of grammar production

rules is obtained for the said non-terminal symbol form the unified grammar specifications. Take each symbol one by one from a list of terminal symbols and/or nonterminal symbols corresponding to the source language grammar, determine whether it is a terminal symbol or a nonterminal symbol. For each terminal symbol obtained which is equivalent to a corresponding symbol in the list of tokens form the source language, consider the next symbol in the list of said terminal symbols and/or nonterminal symbols. For each nonterminal symbol obtained which refers to another non-terminal symbol obtain a set of grammar production rules for that nonterminal symbol and repeat the previous steps.

If all the symbols in the said list of terminal symbols and/or non-terminal symbols corresponding to the source language grammar match with symbols in the said list of tokens of the input text obtain a list of symbols corresponding to the target language grammar from the said unified grammar production rule. For those symbols in the said list of terminal symbols and/or non-terminal symbols which do not match with symbols in the said list of tokens, repeat the earlier steps considering the next production rule from the set of production rules obtained for the non-terminal.

Taking each symbol one by one from the said list of symbols corresponding to the target language grammar, determine whether it is a terminal symbol or non-

terminal symbol. Each terminal symbol obtained are provided as output. For each nonterminal symbol, obtain another unified grammar production rule corresponding to that nonterminal symbol and repeat this step till all the symbols in the said list of symbols corresponding to the target language grammar are exhausted.

BRIEF DESCRIPTION OF DRAWINGS :

Figure I shows a system with which the method according to the invention can be implemented.

Figure II shows the flow chart of the method according to the invention.

Figure III shows the steps taken to create the unified grammar specification in the second step shown in figure II.

Figure IV shows the steps taken to determine if all symbols in a unified grammar production rule match with the symbols in the token list 'T' in the sixth step of figure II and the seventh step of figure V.

Figure V shows the steps taken to determine if a symbol from a unified grammar production rule matches with a symbol from the token list 'T' in the fourth step of figure IV.

Figure VI shows the steps taken to obtain the sentence 'L_t' from a unified grammar production rule 'P' in the eight step of figure II and in the seventh step of figure VI.

DESCRIPTION WITH REFERENCE TO THE DRAWINGS :

The method according to the invention can be implemented by using a processing device (1) such as a microprocessor, a memory (2) and a user input device (3) connected to said processor (1). The user-input device may be a keyboard or any other device which can provide information signals to the processor. The memory typically consists of a RAM and a ROM. According to the invention, the method of automatic translation of a sentences from a source language L_s selected from a number of languages L_1 to L_n to a target language L_t selected from the number of languages L_1 to L_n comprises the following steps.

Step 1 : Grammars G_1 to G_n of all the languages L_1 to L_n respectively and a text 'S' in the source language L_s are provided as inputs.

Step 2 : A unified grammar specification UG is created for the grammars G_1 to G_n.

Step 3 : The input text 'S' in the source language L_s is separated into a list of tokens T using a lexical analyser for the source language L_s.

Step 4 : A nonterminal symbol 'E' is set to the start symbol of the unified grammar specification UG.

Step 5 : A set of grammar production rules P_e is obtained by selecting the production rules which contain 'E' as their target non-terminal symbol from the unified grammar specification UG.

Step 6 : For each unified grammar production rule P in the set of grammar production rules P_e taking each symbol one by one from a list of terminal symbols and/or non-terminal symbols corresponding to the source language grammar G_s, determine whether it is a terminal symbol or a non-terminal symbol.

Step 7 : For each terminal symbol obtained from the previous step which is equivalent to a corresponding symbol in the list of tokens T of the input text in the source language L_s, consider the next symbol in said list of terminal symbols and/or nonterminal symbols corresponding to the source language grammar G_s and for each nonterminal symbol obtained from the previous step which refers to another nonterminal symbol E_s, of the unified grammar specification UG, repeat step (5) onwards with the new nonterminal E_s.

Step 8 : If all the symbols in the said list of terminal symbols and/or non-terminal symbols corresponding to the source language grammar G_s match with all the symbols in the list of tokens T of the input text in the source language L_s, obtain a list of symbols t corresponding to the target language grammar G_t from the unified grammar production rule P and for those symbols which do not match, repeat step 6 onwards for the next unified grammar production rule P defined for the nonterminal symbol 'E'.

Step 9 : Take each symbol one by one, from the list of symbols t corresponding to the target grammar G_t and determine whether it is a terminal symbol or a non-terminal symbol.

Step 10 : For each terminal symbol obtained from the previous step output the symbol, and consider the next symbol and for each nonterminal symbol obtained from the previous step, obtain another unified grammar production rule P corresponding to that nonterminal symbol and repeat the previous step with the new unified grammar production rule, till all the symbols in the list of symbols t corresponding to the target language grammar G_t are exhausted.

The unified grammar specification UG, for the grammars G_1 to G_n of languages L_1 to L_n, is created by defining a unified production rule P_1 in the unified grammar specification UG having the target nonterminal symbol of the production rule P as its target nonterminal symbol for every production rule P of the grammars G_1 to G_n and creating a list of terminal symbols and/or nonterminal symbols in the said production rule P_1 for each grammar G_1 to G_n; adding each and every symbol in the list of terminal and/or nonterminal symbols that are represented by the target nonterminal symbol in the production rule P to the said unified production rule P_1 and repeating previous steps for the next production rule of the grammars G_1 to G_n.

The method according to the invention can be used to represent the equivalence of multiple language grammars and for translating sentences of one language to another.